Synthèse bibliographique : Détection d'émotion dans la voix.

Guillaume Cavelier et Mélanie Dumont

***Abstract*— Ce document constitue une synthèse bibliographique sur la détection des émotions dans la voix et est demandée dans le cadre de l'Élément Constitutif (EC) Interaction Homme-Machines Évoluées dispensé à l'INSA Rouen Normandie. Ce document est à destination de M. Alexandre Pauchet, responsable de l'EC. Dans cette synthèse, nous allons présenter la manière dont peuvent être représentées les émotions humaines afin d'être identifiées par une machine. Nous verrons des exemples de bases de données émotionnelles existantes, certaines caractéristiques qui permettent de déterminer une émotion à partir d'un enregistrement puis nous présenterons les classifieurs utilisés dans ce domaine.**

I. INTRODUCTION

La reconnaissance automatique de l'émotion dans un discours, ou Speech Emotion Recognition (SER) est une discipline ancienne. En effet, le premier article sur ce sujet date de 1996 (*Recognizing emotion in speech*, F. Dellaert). La reconnaissance d'émotion a un intérêt dans le cadre des Interactions Homme-Machine car nous souhaiterions que la machine puisse réagir au ressenti de son interlocuteur. Les assistants vocaux pourraient, par exemple, essayer de calmer une personne énervée. Pour analyser ces émotions, la SER se base sur des techniques de Machine Learning. Nous allons ainsi retrouver un certain nombre d'étapes (présentées en figure 1) qui vont permettre de passer d'un enregistrement audio à l'émotion qu'il véhicule. Nous allons ainsi expliquer ces différentes étapes tout au long de cette synthèse.

II. REPRÉSENTATION DES ÉMOTIONS

Dans le domaine de la SER, il existe deux manières de représenter les émotions[10]. La première est discrète et nous travaille avec sept émotions basées sur le modèle de Ekman. Ces émotions sont la joie, la tristesse, la colère, la peur, le dégoût, la surprise et l'absence d'émotion précise (état neutre). La seconde manière est continue et contient 14 émotions. L'idée de ce modèle est de montrer que les émotions sont des choses ambigus qu'on ne peut pas vraiment classifier en seulement sept états. Ce modèle repose sur deux valeurs : valence et arousal. La valence correspond à une émotion positive ou négative et l'arousal correspond à une forte ou à une basse activité. L'excitation correspond ainsi à une forte valence et une forte arousal (émotion positive et dynamique) alors que la dépression correspond à une basse valence et une basse arousal (émotion négative et amorphe). Ce modèle est détaillé en [7].

III. BASE DE DONNÉES ÉMOTIONNELLES

Une base de données émotionnelle est une base de données qui contient des enregistrements étiquetés selon l'émotion qu'il véhicule. On distingue trois types de base de données émotionnelles [3][9].

Le premier type correspond à des enregistrements joués par des acteurs. Les émotions sont déterminées à l'avance et étiquetées en fonction. Ces bases font parties des plus utilisées dans le domaine de la recherche. D'après certains articles [13], il existe une différence entre les émotions jouées et les émotions réelles. L'utilisation de ce type de base est donc contestée à présent.

Le deuxième type correspond à des enregistrements réels étiquetés par des humains. Ces enregistrements proviennent de call-centers par exemple. Le problème avec ce type de base est que la perception des émotions est différente selon les individus. Une personne qui étiquette un enregistrement peut donc se tromper sur l'émotion réellement exprimée.

Le dernier type de base correspond à des enregistrements réels auto-étiquetés par la personne enregistrée. Les émotions sont réelles et non-soumises à l'appréciation d'une tierce personne. Parmi les bases de données les plus utilisées et disponibles publiquement, nous allons trouver EMODB, une base de données allemande et EMOVO, une base de données italienne.

IV. EXTRACTION DE CARACTÉRISTIQUES

Nous allons distinguer deux types de caractéristiques[5][6][9] présentent dans les enregistrements permettant de détecter l'émotion dans la voix. Le premier type correspond aux caractéristiques prosodiques. Ces caractéristiques vont être la hauteur (Pitch), l'intensité, le débit de parole, l'intonation et la qualité de la voix. La hauteur et l'intensité semblent ainsi liées à l'émotion perçue. Par exemple, une personne en colère va parler fort, vite et va présenter beaucoup d'énergie dans les hautes fréquences.

Le second type correspond aux caractéristiques spectrales du signal. L'une des caractéristiques spectrales les plus utilisées est le Mel Frequency Cepstral Coefficients (MFCC)[4][14]. Une autre caractéristique proposée repose sur l'analyse de la texture du spectrogramme[15]. En fonction de la variation des pixels de l'image du spectogramme, il est possible de déterminer l'émotion correspondant au signal. Les caractéristiques prosodiques sont le plus souvent considérées comme les plus pertinentes dans le cadre de la SER[5]. Il s'avère que la combinaison prosodique/spectrale est celle qui amène aux meilleurs résultats. En pratique, nous allons utiliser des algorithmes qui vont calculer un ensemble de caractéristiques prédéfinies (spectrales et prosodiques) et les passer ensuite à un sélecteur.

V. SÉLECTION DE CARACTÉRISTIQUES

Une fois nos caractéristiques obtenues, il est nécessaire de les filtrer pour ne garder que les plus pertinentes. En effet, les performances des classifieurs sont moins

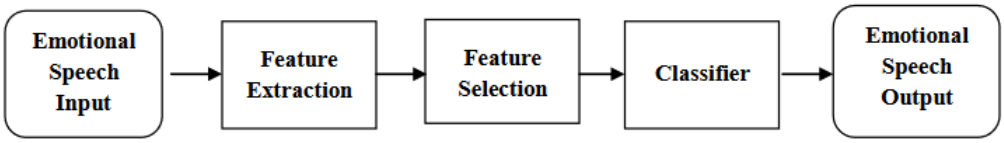

Fig. 1. Chaine de traitement de la Speech Emotion Recognation (SER), d'après [3]

bonnes lorsqu'ils sont soumis à un grand nombre de caractéristiques d'entrées différentes[9]. De plus, cette sélection va également permettre de réduire le temps d'apprentissage[7]. Un algorithme possible pour la sélection de caractéristiques est le Forward Selection[3].

VI. Choix du classifieurs

Une fois la selection de caractéristiques effectuées, il faut choisir un classifieur. Celui-ci va prendre en entrée les caractéristiques extraites et les échantillons de voix, et va regrouper les échantillons qui ont des propriétés similaires. Dans le domaine de la détection d'émotion dans la voix, les classifieurs les plus utilisés sont les suivants :

- Support Vector Machine (SVM)
- Hidden Markov Model (HMM)
- Gaussian Mixtures Model (GMM)
- Artificial Neural Network (ANN)
- k-nearest neighbors (KNN)

A. Support Vector Machine

Le SVM est le résultat d'un calcul d'un ensemble d'algorithmes de machine learning souvent utilisés pour la reconnaissance de pattern. Il obtient généralement de bon résultat mais il faut savoir qu'il est limité en terme de taille des données d'apprentissage [2]. Il est cependant souvent utilisé pour les bonnes performances qu'il obtient [8].

B. Hidden Markov Model

Le HMM est étudié depuis longtemps par les chercheurs pour la reconnaissance d'émotion dans la voix. C'est l'un de classifieurs les plus utilisés. Il est utile pour traiter les aspects statistiques et séquentiels des signaux vocaux. Cependant, sa capacité de classifieur n'est pas satisfaisante en elle-même. [3] C'est pourquoi il est souvent associé à un autre classifieur comme le GMM.

C. Gaussian Mixtures Model

Le GMM est utilisé quand les caractéristiques globales sont extraites d'énoncés d'entraînement (training speech) [1]. Les équations d'apprentissages et de tests sont basées sur la supposition que tous les vecteurs sont indépendants. Il est utile sur des modèles utilisant des caractéristiques spectrales. C'est aussi la méthode statistique la plus affinée pour l'estimation de la densité et le clustering [3].

D. Artificial Neural Network

L'ANN est connue pour être plus efficace pour les modélisations non-linéaire. De plus, ces résultats peuvent être meilleurs que le HMM ou le GMM quand il y a peu de données d'apprentissage. [3]

E. k-nearest neighbors

Le KNN est le classifieur le plus traditionnel pour la reconnaissance statistique supervisée de pattern [2].

F. Comparaison des classifieurs

L'article "Speech Emotion Recognition" paru dans le "International Journal of Soft Computing and Engineering (IJSCE)" en 2012 fait une comparaison des différents classifieurs mentionnés. Il donne le pourcentage d'exactitude pour un système dépendant du locuteur et un système indépendant du locuteur pour chaque classifieur. Voici un tableau récapitulant les résultats obtenues [1] :

Classifieur	Reconnaissance indépendant du locuteur	Reconnaissance dépendant du locuteur
HMM	64,77%	76,12%
SVM	75%	80%
ANN	52,87%	51,19%
GMM	75%	89,12%
KNN	N/A	N/A

Nous pouvons remarquer que le GMM et le SVM obtienne de très bon résultat et que la reconnaissance dépendant du locuteur obtient généralement de meilleurs résultats.

Des études plus récentes ont également montrées une grande progression pour les réseaux de neurones à convolution, certains modèles atteignant plus de 90% de détections positives[11].

Chaque classifieur à des avantages et des limites. Ainsi, certains chercheurs se sont penchés sur la possibilités de créer des modèles qui combineraient des classifieurs pour ainsi tirer le meilleurs de chacun d'entre eux. Une équipe du laboratoire de Shanghai a mis au point un système hybride basé sur le SVM et un modèle HMM-GMM [12]. D'après eux, le nouveau système obtient des résultats bien meilleurs que sur les classifieurs indépendants. De plus, le système semble bien prendre en compte les avantages des différents classifieurs.

VII. CONCLUSIONS

La Speech Emotion Recognition reste une discipline qui, bien qu'âgée de 22 ans, attire toujours autant les chercheurs. Comme présenté plus haut, les résultats obtenus grâce au Machine Learning peuvent encore être améliorés. Il existe différentes pistes d'amélioration pour cette discipline [10]. Nous pouvons notamment citer le Holistic Speaker Modeling qui a pour but de tenir compte de l'état du locuteur (enrhumé, alcoolique, ...) pour déterminer son émotion, ou l'amélioration des bases de données déjà existantes qui pourraient de plus en plus se tourner vers des données non jouées. La démocratisation des réseaux de neurones devrait également permettre d'améliorer les résultats actuels[11]. De plus, l'utilisation du Deep Learning permettrait également de s'affranchir de l'étape d'extraction des caractéristiques.

REFERENCES

[1] Ashish B Ingale and D S Chaudhari. Speech Emotion Recognition. 2(1) :4, 2012.

[2] Aastha Joshi and Rajneet Kaur. A Study of Speech Emotion Recognition Methods. (4) :4, 2013.

[3] Dipti D Joshi and M B Zalte. Speech Emotion Recognition : A Review. page 4.

[4] S. Lalitha, D. Geyasruti, R. Narayanan, and Shravani M. Emotion Detection Using MFCC and Cepstrum Features. *Procedia Computer Science*, 70 :29–35, 2015.

[5] Iker Luengo, Eva Navas, and Inmaculada Hernaez. Feature Analysis and Evaluation for Automatic Emotion Identification in Speech. *IEEE Transactions on Multimedia*, 12(6) :490–501, October 2010.

[6] Marko Lugger and Bin Yang. On the Relevance of High-Level Features for Speaker Independent Emotion Recognition of Spontaneous Speech. page 4.

[7] Arianna Mencattini, Eugenio Martinelli, Giovanni Costantini, Massimiliano Todisco, Barbara Basile, Marco Bozzali, and Corrado Di Natale. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. *Knowledge-Based Systems*, 63 :68–81, June 2014.

[8] Yixiong Pan, Peipei Shen, and Liping Shen. Speech Emotion Recognition Using Support Vector Machine. *International Journal of Smart Home*, 6(2) :8, 2012.

[9] S. Ramakrishnan. Recognition of Emotion from Speech : A Review. In S Ramakrishnan, editor, *Speech Enhancement, Modeling and Recognition- Algorithms and Applications*. InTech, March 2012.

[10] Björn W. Schuller. Speech emotion recognition : two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5) :90–99, April 2018.

[11] Somayeh Shahsavarani. Speech Emotion Recognition using Convolutional Neural Networks. page 87.

[12] Kaiyu Shi, Xuan Liu, and Yanmin Qian. Speech Emotion Recognition Based on SVM and GMM-HMM Hybrid System. page 5.

[13] T. Vogt and E. Andre. Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. In *2005 IEEE International Conference on Multimedia and Expo*, pages 474–477, Amsterdam, The Netherlands, 2005. IEEE.

[14] Siqing Wu, Tiago H. Falk, and Wai-Yip Chan. Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53(5) :768–785, May 2011.

[15] Turgut Özseven. Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition. *Applied Acoustics*.